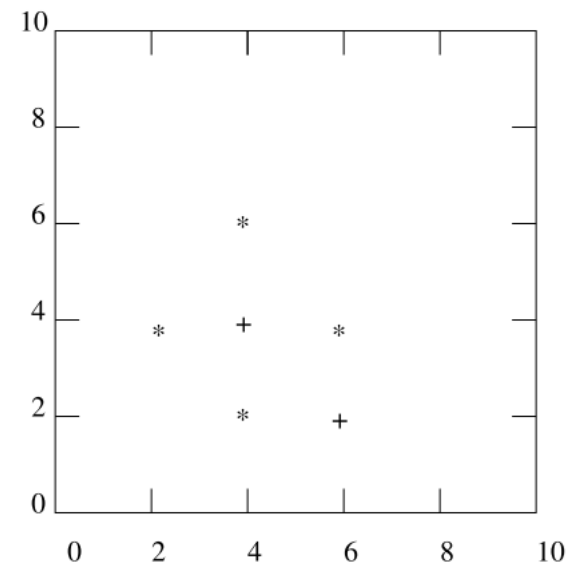


Bài tập k-NN

Phạm Nguyên Khang
pnkhang@cit.ctu.edu.vn

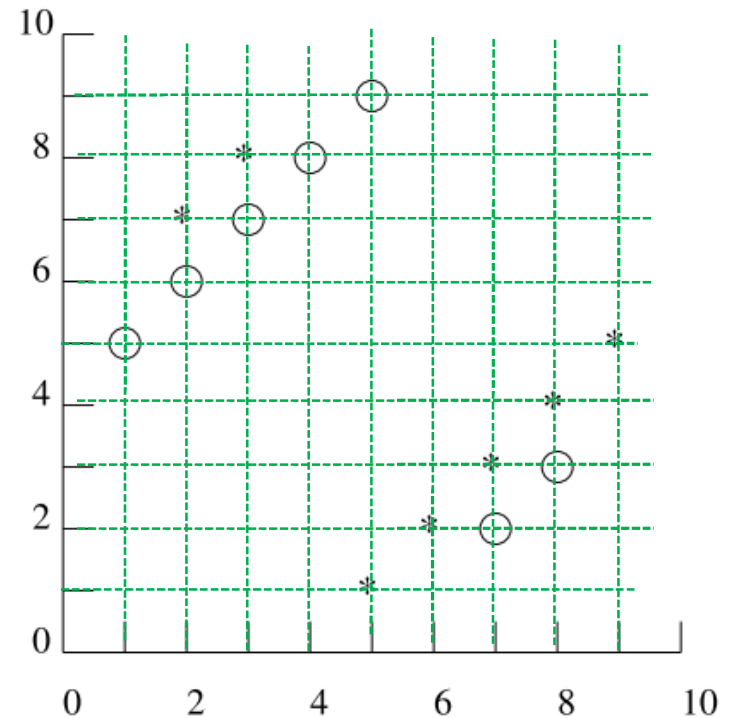
Sử dụng bộ phân lớp k-NN

- Cho tập dữ liệu học như hình bên
 - a) Vẽ đường biên quyết định cho giải thuật 1-NN giả sử ta sử dụng khoảng cách Euclid.
 - Dấu cộng (+) thuộc lớp dương.
 - Dấu sao (*) thuộc lớp âm.
 - b) Điểm (8, 1) sẽ được phân vào lớp nào nếu sử dụng bộ phân lớp 1-NN ?
 - c) Điểm (8, 8) sẽ được phân vào lớp nào nếu sử dụng bộ phân lớp 1-NN ?



Xác định giá trị k

- Cho tập dữ liệu như hình bên
 - a. Tìm k để lỗi trên tập huấn luyện nhỏ nhất, và giá trị nhỏ nhất đó bằng bao nhiêu ? (leave-one-out)
 - b. Tại sao lỗi trên tập huấn luyện không phải là một ước lượng hợp lý cho lỗi trên tập kiểm tra (test), cụ thể trong trường hợp trên là giá trị k được ước lượng từ câu a.
 - c. Tại sao sử dụng giá trị k quá lớn sẽ cho kết quả không tốt trong tập dữ liệu này ?
 - d. Tại sao sử dụng giá trị k quá nhỏ sẽ cho kết quả không tốt trong tập dữ liệu này ?



Chứng minh lý thuyết (1/2)

- Một kết quả nổi tiếng của Cover và Hart (1967) tỉ lệ lỗi của bộ phân lớp 1-NN nhiều nhất là 2 lần tỉ lệ tối ưu Bayes.
- Cho tập $X = \{x_1, x_2, \dots, x_m\}$ là tập huấn luyện trong không gian R^d , gọi:
 - $y_1, y_2, \dots, y_m$ ($y_i \in Y = \{0, 1\}$) là nhãn của các phần tử.
 - $p_i(x) = p(X = x \mid Y = i)$ phân phối xác suất có điều kiện của X khi biết Y . Giả sử $p(x) > 0$ với mọi x .
 - Đặt $R = p(Y = 1)$ xác suất một điểm thuộc lớp 1, giả sử $0 < R < 1$.

Chứng minh lý thuyết (2/2)

- Hãy tính xác suất: $q(x) = p(Y = 1 \mid X = x)$
 - (Sử dụng định lý Bayes)
- Biểu diễn $q(x)$ theo $p_0(x)$, $p_1(x)$, và R .
- CMR: $0 < q(x) < 1$.
- Bộ phân lớp tối ưu Bayes là một bộ phân lớp mà nó luôn gán một phần tử x vào lớp i có xác suất $P(Y = i \mid x)$ lớn nhất.
 - $\text{label}(x) = \arg\max P(Y=i \mid x)$ với $i = 0$ hoặc 1 .
 - Lỗi phân lớp là $1 - P(Y = \text{label}(x) \mid x)$
- Cho một phần tử kiểm tra, x viết công thức tính lỗi của bộ phân lớp tối ưu Bayes theo $q(x)$?